12-2001

Dimension and exercise variance in assessment center scores: A large-scale evaluation of multitrait-multimethod studies

Filip LIEVENS
Singapore Management University, filiplievens@smu.edu.sg

James M. CONWAY
Ghent University
DOI: https://doi.org/10.1037/0021-9010.86.6.1202

Published in Journal of Applied Psychology, 2001
December, Volume 86, Issue 6, Pages 1202-1222
http://doi.org/10.1037//0021-9010.86.6.1202

Dimension and Exercise Variance in Assessment Center Scores: A Large-Scale Evaluation of Multitrait–Multimethod Studies

Filip Lievens
Ghent University

James M. Conway
Central Connecticut State University

This study addresses 3 questions regarding assessment center construct validity: (a) Are assessment center ratings best thought of as reflecting dimension constructs (dimension model), exercises (exercise model), or a combination? (b) To what extent do dimensions or exercises account for variance? (c) Which design characteristics increase dimension variance? To this end, a large set of multitrait–multimethod studies ($N = 34$) were analyzed, showing that assessment center ratings were best represented (i.e., in terms of fit and admissible solutions) by a model with correlated dimensions and exercises specified as correlated uniquenesses. In this model, dimension variance equals exercise variance. Significantly more dimension variance was found when fewer dimensions were used and when assessors were psychologists. Use of behavioral checklists, a lower dimension–exercise ratio, and similar exercises also increased dimension variance.

Although personnel selection has generally been regarded as an applied area with a heavy emphasis on predictive efficiency, the construct-driven approach has in recent years gained in importance among researchers and practitioners (Binning & Barrett, 1989; Klimoski, 1993; Schmitt & Chan, 1998). A tenet of this construct approach is that whereas high predictive validity of selection practices is certainly desirable, it is also crucial to understand why selection devices work and what exact constructs they measure. The need for construct validity is even clearer when feedback for personal development supplements the selection decision: If invalid measures are used, feedback will be inappropriate and potentially detrimental.

The assessment center (AC) is one of the personnel measurement instruments, used for both employee selection and personal development, for which construct validity has come under scrutiny in recent years. ACs were originally conceived as a way to measure stable individual-differences attributes, such as planning, problem solving, or sociability (Sackett & Dreher, 1982). These attributes, usually referred to as dimensions, were seen as the building blocks of ACs. The assumed underlying model was therefore a *dimension-based model.*

However, early confirmatory factor analysis (CFA) studies on the construct validity of ACs (e.g., Bycio, Alvares, & Hahn, 1987) concluded that the dimensions did not emerge as important factors. Instead, the AC ratings combined into exercise factors. From a conceptual standpoint, these prior findings imply that the dimensions are not essential building blocks of ACs, prompting the conclusion that they should be eliminated from the AC framework (Lowry, 1997; Robertson, Gratton, & Sharpley, 1987). An AC is then reconceptualized as a series of miniaturized work samples designed to elicit job-relevant managerial behavior (Robertson et al., 1987; Sackett & Dreher, 1982). This can be referred to as the *exercise-based model* of AC ratings.

More recent CFA studies (e.g., Donahue, Truxillo, Cornwell, & Gerrity, 1997; Kudisch, Ladd, & Dobbins, 1997) found that both exercises and dimensions emerged as important factors. In this *combination model,* exercise variance was also found to dominate dimension variance. According to such recent studies, the evaluation of dimension and exercise effects is also complicated because no two ACs are alike and the quality of dimension measurement in ACs may be largely dependent on the design of the AC under investigation—the predominance of exercise versus dimension effects may depend on specific AC design characteristics.

It can be argued that prominent exercise effects (either in the exercise-based model or in the combination model) are not necessarily troublesome for selection applications of ACs because the multiple simulation exercises contribute to the content and predictive validity of ACs. However, a theoretical understanding of what is measured in ACs should facilitate efforts to improve validity in the future. More immediately, knowing whether AC ratings reflect dimensions or exercises has important implications for the way a job analysis should be conducted (i.e., whether knowledge, skills, and abilities or tasks should be the focus of the job analysis conducted).

For the developmental use of ACs more so than for selection, prominent exercise effects have disturbing implications. Developmental feedback is based on the dimension-based model, so if the dimensions are not valid indicants of the managerial abilities, the feedback and subsequent action plans could have detrimental

Filip Lievens, Department of Personnel Management and Work and Organizational Psychology, Ghent University, Ghent, Belgium; James M. Conway, Department of Psychology, Central Connecticut State University.

Parts of this article were presented at the 15th Annual Conference of the Society for Industrial and Organizational Psychology, April 2000, New Orleans, Louisiana. Filip Lievens is a postdoctoral research fellow of the Fund of Scientific Research-Flanders (Belgium) (F.W.O.).

We acknowledge the suggestions of Wilfried De Corte, David Kenny, Richard Klimoski, and Charles Lance on earlier versions of this article.

Correspondence concerning this article should be addressed to Filip Lievens, Department of Personnel Management and Work and Organizational Psychology, Ghent University, Henri Dunantlaan 2, 9000 Ghent, Belgium. Electronic mail may be sent to filip.lievens@rug.ac.be.

effects (Bycio et al., 1987; Fleenor, 1996; Joyce, Thayer, & Pond, 1994). The following example by Kudisch et al. (1997) exemplifies these practical ramifications:

> Telling a candidate that he or she needs to improve his or her overall leadership skills may be inappropriate if the underlying construct being measured is dealing with a subordinate in a one-on-one situation (i.e., tapping individual leadership as opposed to group leadership). (p. 131)

In other words, although ACs for selection purposes may not require the dimension-based model, ACs purported to identify and develop managerial strengths and weaknesses do require that the dimensions emerge as (the most) important factors.

On the basis of the issues raised in the preceding paragraphs, this study addresses the following three questions: (a) Which model (a dimensions-only model, an exercises-only model, or a combination model) serves as the best underlying conceptualization of AC ratings? (b) If ACs reflect a combination of dimension and exercise effects, what is the relative importance of the individual-differences constructs (dimensions) vis-à-vis the exercises? In other words, to what extent do dimensions or exercises account for variance? (c) Which specific AC design characteristics increase the quality of dimension measurement (i.e., increase proportions of dimension variance) in ACs? Although in the past single studies tried to answer these questions, the central premise of this study is that only a large-scale systematic investigation of many AC studies (i.e., a large set of multitrait–multimethod [MTMM] matrices) may provide a more definite answer to them.

A Closer Look at Dimension and Exercise Effects in ACs

As we noted above, the traditional assumption underlying ACs is that AC ratings are primarily a function of the assessee's standing on a dimension and that the assessee's performance on the dimensions will be relatively stable across exercises. Exercises are then viewed as nothing more than alternative ways to measure the same traits. Translated to correlations in an MTMM matrix (Campbell & Fiske, 1959), this view implies that ratings on the same dimension across exercises will be particularly consistent and will correlate substantially. In factor analysis terms, this view assumes that AC ratings will primarily reflect dimension factors.

By way of an example, Table A1 presents the results of an MTMM matrix of AC ratings (Arthur, Woehr, & Maldegen, 2000), which lends good support to these assumptions. The summary correlations of the MTMM matrix show that dimensions were consistently measured across the exercises because the average same-dimension correlations equal .60. This matrix also shows a lack of exercise effects, with different-exercise correlations almost identical to same-exercise correlations (M = .33 vs. .39, respectively). As we discuss later in this article, results of structural equation models applied to this MTMM matrix (Table A2, Table A3, and Table A4) reveal similar conclusions about the dimension and exercise effects present in this MTMM matrix of AC ratings because dimension factor loadings are virtually always substantially larger than exercise loadings. In short, the MTMM and structural equation analyses support the use of dimensions for this AC, and the developmental use of dimension scores seems appropriate.

As we already mentioned, another view is that the simulation exercises are the cornerstones of ACs and that AC ratings are primarily a function of the assessee's standing on an exercise, regardless of the dimensions. Translated to correlations in an MTMM matrix, this view implies that ratings on various dimensions within an exercise will be particularly consistent and will correlate substantially. In factor analysis terms, this view assumes that AC ratings will mainly reflect exercise factors.

By way of an example, Table B1 presents an MTMM matrix of AC ratings (Schneider & Schmitt, 1992), which lends support to the aforementioned assumptions. This MTMM matrix shows little evidence of dimension effects. For instance, ratings on the same dimension across exercises were not particularly consistent because the mean same-dimension correlation (.25) was much smaller than the mean same-exercise correlation (.72). This very high mean same-exercise correlation, which indicates that the ratings on different dimensions cluster within single exercises, is indicative of strong exercise effects. As we discuss later in this article, similar conclusions about the dimension and exercise effects present in this AC MTMM matrix are derived when structural equation modeling (see Table B2, Table B3, and Table B4) is used because exercise factor loadings are almost always substantially larger than dimension factor loadings. Hence, the MTMM and CFA analyses provide little support for the use of dimensions as building blocks for this AC, and use of dimension scores for feedback and action planning would be misleading and inappropriate.

It is clear that these two matrices were chosen for illustration purposes because they represent relative extremes. However, our study asks which model—dimensions-only, exercises-only, or a combination—is typical for AC MTMM matrices. In addition, we examine what the relative sizes of dimension and exercise effects are. Because the answers to this question will probably depend on the AC under consideration, our study also seeks to explain variation across MTMM matrices by asking what AC design characteristics are associated with more dimension variance.

Design Characteristics That Can Increase Dimension Variance

Because of the desirability for ACs to measure individual-differences constructs, it is pivotal to identify design considerations that improve the quality of construct measurement (i.e., increase proportions of dimension variance). It is equally important that these interventions are grounded by theory (Landy, Shankster, & Kohler, 1994). Along these lines, the following three perspectives seem most relevant.

Limited Cognitive Capacity Perspective

The first theoretical perspective is that assessors possess limited information-processing capacities and, therefore, are not always able to meet the cognitive demands of the AC process (Reilly, Henry, & Smither, 1990). For instance, cognitive overload may stem from the fact that the behavioral information is presented to assessors at a very fast rate in the various exercises, which often last longer than 30 min. Cognitive overload may also come from the many inferential leaps that assessors make in order to provide dimensional ratings. The determination of relevance, dimensionality, and relative weight of behaviors is among the inferences typically required of assessors. Because the general assumption is that poor construct measurement results from assessors' inability

to deal with their cognitively complex task, procedural interventions should be proposed to reduce cognitive overload.

One of these interventions may consist of limiting the number of dimensions rated per exercise. Primary studies have supported the viability of this intervention, because notable improvements in the quality of construct measurement were reported when assessors rated fewer dimensions (Gaugler & Thornton, 1989; Maher, 1990) and conceptually distinctive dimensions (Kleinmann, Exler, Kuptsch, & Köller, 1995).

Behavioral checklists, which list per dimension and per exercise the relevant behaviors, may serve as another vehicle to reduce the number of inferences required from assessors. In fact, when using behavioral checklists, assessors are not required to categorize behavior. Instead, they can concentrate their efforts on the observation of relevant behaviors. As argued by Reilly et al. (1990), the checklists may further reduce cognitive demands by serving as retrieval cues to guide the recall of observed behaviors. The research evidence regarding the effectiveness of behavioral checklists, however, is mixed. Some studies have found significant increases in convergent and discriminant validity (Reilly et al., 1990), other studies have reported increases only in discriminant validity (Donahue et al., 1997), and still other research has revealed no real benefits (Sweeney, 1976).

A final (indirect) intervention to help assessors may involve making the dimensions transparent to assessees. In several studies, Kleinmann and colleagues (Kleinmann, 1997; Kleinmann & Köller, 1997; Kleinmann, Kuptsch, & Köller, 1996) found that in this case assessees oriented themselves toward the given dimensions and demonstrated more clearly and consistently the accompanying behaviors. Because this intervention increases the opportunity for assessors to observe dimension-relevant behavior, the assumption is that, in turn, their task becomes less complex, resulting in better construct measurement. Kleinmann and colleagues (Kleinmann, 1997; Kleinmann & Köller, 1997; Kleinmann et al., 1996) indeed discovered higher dimension variance in the ratings of the group of assessors who observed and rated "informed" assessees. In sum, on the basis of the limited cognitive capacity perspective and the results of primary studies, we hypothesized that significantly higher proportions of dimension variance would be found when fewer dimensions were used, when the dimension–exercise ratio was low, when behavioral checklists were used, and when dimensions were made transparent to assessees.

Expert Assessor Perspective

Differences between experts and novices (Chi, Glaser, & Farr, 1988) may serve as a second theoretical framework to underpin AC design interventions. Expert assessors are expected to possess and use well-established cognitive structures when they are rating assessees. For expert assessors, these organizing frameworks are helpful because they guide attention, categorization, integration, and recall processes (Fiske & Taylor, 1991; Zedeck, 1986). In contrast, novice assessors are not expected to possess such well-established cognitive structures when they are rating assessees (Cardy, Bernardin, Abbott, Senderak, & Taylor, 1987).

Because expert assessors typically differ from novice assessors by their education, experience, and training, each of these aspects may guide AC design principles. For instance, higher proportions of dimension variance may be expected when more experienced and better trained people serve as assessors.

In line with this expert assessor perspective, previous primary studies have revealed higher proportions of dimension variance for assessors with a degree in psychology (Lievens, in press; Sagie & Magnezy, 1997) and for assessors who had gained experience through participation in numerous ACs (Lorenzo, 1984). To be specific, Sagie and Magnezy (1997) hypothesized that because of differing prior experiences and educational backgrounds, managers and psychologists use different schemas when rating candidates. As expected, Sagie and Magnezy reported that in the ratings of psychologists, all five predetermined dimensions were represented. Managers' ratings, however, yielded only two dimension factors. Lievens (in press) found that even industrial–organizational psychology students outperformed managers in terms of providing distinct ratings on different dimensions.

The effects of training have been more equivocal. Some studies have reported higher quality of construct measurement among assessors who followed a more comprehensive (frame-of-reference) training program (Lievens, 2001; Schleicher, Day, Mayes, & Riggio, 1999). Furthermore, Lorenzo (1984) reported lower dimensional accuracy for assessors who had attended assessor training and had been serving as full-time assessors for at least 3 months (compared with novice assessors). Other studies have found no substantial effects of increasing the length of the training given to assessors (Dugan, 1988; Maher, 1995). In short, on the basis of the expert assessor perspective and the results of prior primary studies, we hypothesized higher proportions of dimension variance for psychologist assessors and for assessors receiving a more comprehensive (i.e., longer) assessor training program.

Interactionist Perspective

The third perspective posits that the low convergence between ratings of the same dimension across different exercises (i.e., low convergent validity) reflects real performance differences of candidates. In other words, because candidates often perform in various different exercises, some candidates may actually perform better on the same dimensions in some exercises (e.g., individual exercises) than in other exercises (e.g., group exercises), resulting in a lack of cross-situational consistency (Neidig & Neidig, 1984). This perspective builds on the interactionist perspective in social and personality psychology (Magnusson, 1982). According to this perspective, an assessee's behavior is determined neither by the person (assessee) nor by the situation (simulation exercise) but by the interaction of the person and the situation.

Primary studies have used this perspective to examine the nature of the exercises used in ACs. For example, Highhouse and Harris (1993) asked experienced assessors to depict AC exercises in terms of performance constructs elicited (e.g., generates enthusiasm, asks questions). They discovered some evidence for the hypothesis that assessees would be rated more consistently in exercises that were perceived to be more similar. In another study, Schneider and Schmitt (1992) used an experimental design to investigate the impact of two exercise factors (i.e., exercise form and exercise content). They found that exercise form (group discussion vs. role-playing) considerably affected the proportions of variance, with less similar exercises

resulting in more exercise variance and less dimension variance. Exercise content (competitive vs. cooperative) yielded virtually no effects. Taken together, on the basis of the interactionist perspective and the results of prior primary studies, we hypothesized significantly higher proportions of dimension variance for more similar exercises.

To our knowledge, one prior review (Lievens, 1998) attempted to integrate and summarize the effectiveness of different AC design considerations in improving construct validity evidence. However, this was a narrative review, making it difficult to compare the relative value of different AC design interventions. Another shortcoming is that the conclusions regarding the effectiveness of each design intervention were derived from only one or two studies. The statistical approach used to examine construct validity also considerably varied across the studies reviewed.

To alleviate these limitations, in this study we coded AC design characteristics for a large set of MTMM matrices and used statistical tests to examine whether they affected the proportions of dimension and exercise variance. The design characteristics coded across the MTMM matrices are shown as column headers in Table 1.

Method

Literature Search and Criteria for Inclusion

To find MTMM matrices of AC ratings, we conducted a search using a number of computerized databases (i.e., PsycLIT, Dissertation Abstracts International, and Current Contents). Key words included *assessment center* in combination with *construct validity* or *multitrait–multimethod.* In addition, we scrutinized reference lists from obtained studies to find other published and unpublished studies.

The following criteria for including an MTMM matrix were used. First, the so-called within-exercise dimension ratings (i.e., ratings made on completion of each exercise) had to be cast as an MTMM correlation or covariance matrix in which various dimensions served as traits and various AC exercises as methods. Matrices in which the methods represented different assessors or different rating sources were excluded. When studies provided only summary MTMM correlations, the authors were contacted to retrieve the full correlation matrix. However, the full correlation matrices of some earlier studies (e.g., Sackett & Dreher, 1982) were no longer available. Second, the MTMM matrix had to include at least three dimensions and two exercises (see Conway, 1996, for a detailed discussion of this criterion in MTMM matrices). Third, consistent with AC practice, only matrices were included in which assessors rotated across the AC exercises.

Table 1
Description of Multitrait–Multimethod Matrices of Assessment Center Ratings Included in Study

Reference	*N*	No. of dimensions	Dimension–exercise ratio	Behavioral checklists	Transparent dimensions	Type of assessor	Training length	Exercise similarity
Arthur et al. (2000)	149	4	1.33	Not used	No	Psychologist	≤1 day	Dissimilar
A. S. Becker (1990)	81	5	1.25	Not used	No	Manager	>1 day	Dissimilar
A. S. Becker (1990)	81	5	1.25	Not used	No	Manager	>1 day	Dissimilar
Bobrow & Leonards (1997)	196[a]	11 (9)[b]	3.67	Used	No	Psychologist	>1 day	Dissimilar
Bycio et al. (1987)	1,170	8	1.60	Not used	No	Manager	—	Dissimilar
Chorvat (1994)	207	11	2.75	Used	No	—	>1 day	Dissimilar
Donahue et al. (1997)	188	9	2.25	Not used	No	Manager	≤1 day	Similar
Fleenor (1996)	102[a]	10 (5)[b]	1.25	Used	No	—	>1 day	Dissimilar
Fredricks (1989)	66	8 (5)[b]	2.67	Not used	No	Manager	>1 day	Similar
Harris et al. (1993)	793	7	1.40	Not used	No	Manager	>1 day	Dissimilar
Joyce et al. (1994)	75	7 (4)[b]	1.75	Used	No	Manager	>1 day	Dissimilar
Joyce et al. (1994)	77	7 (5)[b]	2.33	Used	No	Manager	>1 day	Dissimilar
Kleinmann et al. (1994)	60	3	1.00	Used	No	Psychologist	≤1 day	Similar
Kleinmann et al. (1994)	60	3	1.00	Used	No	Psychologist	≤1 day	Similar
Kleinmann et al. (1996)	59	3	1.00	Used	No	Psychologist	≤1 day	Dissimilar
Kleinmann (1997)	70	3	1.00	Not used	Yes	Psychologist	≤1 day	Dissimilar
Kleinmann (1997)	62	3	1.50	Used	No	Psychologist	≤1 day	Dissimilar
Kleinmann (1997)	63	3	1.50	Used	Yes	Psychologist	≤1 day	Dissimilar
Kolk et al. (2000)	99	3	1.50	Not used	No	Psychologist	—	Similar
Kolk et al. (2000)	99	3	1.50	Not used	Yes	Psychologist	—	Similar
Kudisch et al. (1997)	138[a]	7	1.75	Not used	No	Psychologist	>1 day	Dissimilar
Lievens & Van Keer (1999)	191	5	0.83	Used	Yes	Psychologist	>1 day	Dissimilar
Parker (1992)	379	11	3.67	—	No	Manager	—	Dissimilar
Robie et al. (2000)	100	4	2.00	Used	No	Psychologist	>1 day	Similar
Sagie & Magnezy (1997)	336[a]	5	1.67	Not used	No	Manager	>1 day	Dissimilar
Sagie & Magnezy (1997)	374[a]	5	1.67	Not used	No	Psychologist	>1 day	Dissimilar
Schleicher et al. (1999)	63	3	1.00	Used	No	Psychologist	≤1 day	Dissimilar
Schleicher et al. (1999)	63	3	1.00	Used	No	Psychologist	≤1 day	Dissimilar
Schneider & Schmitt (1992)	89	3	0.75	Used	No	Manager	>1 day	Dissimilar
Sweeney (1976)	186	3	1.50	Used	No	Manager	>1 day	Similar
Sweeney (1976)	186	3	1.50	Not used	No	Manager	>1 day	Similar
Van der Velde et al. (1994)	88	10 (6)[b]	3.33	Not used	No	Manager	—	Dissimilar
Veldman (1994)	188	6	3.00	—	No	—	—	Similar
Veldman (1994)	71	6	3.00	—	—	—	—	Similar

Note. Dashes indicate that this information was not available.
[a] Because this matrix consisted of correlations computed on a different number of cases, we used the harmonic mean of the *N*s for each individual correlation as input for the analyses (Viswesvaran & Ones, 1995). [b] This parenthetical value refers to the number of dimensions used in the analyses. As noted, we dropped all measures of randomly chosen dimensions so that the sample size was always greater than or equal to the number of parameters estimated.

Fourth, the matrix had to be positive definite. Finally, the sample size had to be greater than 50 and greater than the number of parameters estimated in any of the models (Tanaka, 1987). If the sample size was not greater than the number of parameters estimated in any of the models, all measures of randomly chosen dimensions were removed until the criterion was satisfied. This was done for 6 matrices.

Thirty-four matrices from studies dating from 1976 to 2000 conformed to these criteria. Table 1 presents these studies together with their characteristics. Twenty matrices came from published articles, 9 were unpublished dissertations, and 5 were conference presentations. The mean sample size was 182.62 (*Mdn* = 99, total N = 6,209), the mean number of dimensions was 5.59 (*Mdn* = 5), and the mean number of exercises was 3.29 (*Mdn* = 3). Twenty matrices (59%) came from field studies; the rest were lab studies. In 21 (62%) of the 34 matrices, a fully crossed design was used, which means that each dimension was measured in each exercise.

Structural Equation Models

Most early studies on AC construct validity used Campbell and Fiske's (1959) MTMM approach to evaluate dimension and exercise effects. The most important limitation of this approach is that it does not yield any precise criterion for evaluating which model is most appropriate or for estimating the size of dimension and exercise effects (Schmitt & Stults, 1986; Widaman, 1985). Hence, in recent years, AC researchers have used structural equation modeling (i.e., CFA) as a more powerful analytic tool to evaluate the MTMM matrix. The MTMM matrix is then explained in terms of underlying constructs rather than observed variables. Furthermore, fit indices and parameter estimates of the models provide information on the convergent and discriminant validity as well as on the dimension and exercise variance present in the AC. Competing models can also be statistically compared.

Consistent with our first research question, we tested a comprehensive set of structural equation models that represented different conceptualizations of ACs. First, we tested a dimensions-only model, including a factor for each AC dimension (the dimension factors were allowed to be correlated) but ignoring exercises. That is, in the first model, there was no attempt to estimate exercise effects. The dimensions-only model did include random error variance, specified as unique factors for each variable. Second, we tested an exercises-only model. This model can be thought of as the opposite of the dimensions-only model—it included a factor for each AC exercise (with exercise factors allowed to be correlated) but ignored dimensions. Again, each variable had a unique factor. Third, we tested combination models. The estimation of combination models allowed us to see if including both dimension and exercise effects resulted in a better fit than including only one or the other.

A complication with combination models is that there are several possible such models, and there has been controversy over which model to use (e.g., Lance, Noble, & Scullen, 2000). We therefore estimated and compared a number of combination models, which are described below. In particular, we followed Conway's (1996) strategy for analyzing MTMM data. Conway recommended that researchers begin with the general CFA model (including correlated dimension factors and correlated exercise factors), followed by the direct product model and the correlated uniqueness model. The following paragraphs briefly describe each of the combination models that were tested. To help in understanding these models, we provide each model's standardized parameter estimates for the MTMM matrices in Appendix A and Appendix B. Standardized estimates are particularly useful because squared dimension and exercise factor loadings can be interpreted as proportions of dimension and exercise variance.

General CFA model. In this model, the correlations among variables are an additive function of both correlated dimension factors and correlated exercise factors along with unique factors (Schmitt & Stults, 1986; Widaman, 1985). We also estimated a variation of the general CFA model that includes correlated exercise factors but only one global dimension factor. This model represents the assumption that assessors are unable to distinguish among the various dimensions (see also Bycio et al., 1987; Kudisch et al., 1997).

A CFA model with correlated dimensions and correlated exercises yields estimates of dimension loadings, exercise loadings, uniquenesses, dimension factor correlations, and exercise factor correlations, as illustrated by Table A2 and Table B2. The squared standardized dimension loadings represent proportions of dimension variance, and the squared standardized exercise loadings reflect proportions of exercise variance (Widaman, 1985). For instance, in Table A2, the mean proportion of dimension variance across all measures is .49, and the mean proportion of exercise variance is .21. Such results indicate that dimensions are more important factors than exercises and generally support the interpretation of AC ratings in terms of the intended dimensions. In contrast, the CFA results in Table B2 indicate that exercise variance (i.e., mean of squared exercise loadings = .49) dominates dimension variance (i.e., mean of squared dimension loadings = .29) and provide less support for the dimensions as building blocks of ACs.

Although the general CFA model has been used in previous AC research, an important drawback is that recent large-scale studies in the MTMM literature have shown serious estimation problems for this model. A common example is improper parameter estimates (outside the permissible range; e.g., factor correlations > 1 or negative variances; T. E. Becker & Cote, 1994; Brannick & Spector, 1990; Conway, 1996; Marsh & Bailey, 1991). More important, AC studies have been plagued by the same problems (Bycio et al., 1987; Kudisch et al., 1997; Lievens & Van Keer, 1999; Van der Velde, Born, & Hofkes, 1994). Table B2 shows that such improper estimates were also found for the MTMM matrix of Schneider and Schmitt (1992). In particular, one of the exercise factor correlations was held at a boundary value of −1.00. Another potential problem was raised by Marsh (1989), who presented evidence that the general CFA model inflates method (e.g., exercise) variance estimates.[1] If this has happened in AC research, then the earlier findings of exercise predominance may be an artifact of the analysis model. Because of estimation problems and the possibility of overestimation of exercise variance, estimates based on this analytic model are questionable. Hence, researchers are urged to test alternatives to this correlated dimensions and correlated exercises CFA model (Conway, 1996).

Direct product model. The direct product model was proposed by Browne (1984; Wothke & Browne, 1990). This model is mathematically related to early research on three-mode factor analysis (Levin, 1965; Tucker, 1966) and has its conceptual roots in Campbell and O'Connell's (1967, 1982) research on multiplicative MTMM effects. Campbell and O'Connell (1967, 1982) suggested that the common assumption that method effects combine additively with trait effects to inflate all same-method correlations by about the same amount may be wrong. Instead, they proposed that traits and methods combine multiplicatively, affecting high correlations more than low ones. Furthermore, the impact of method effects can be to either augment or attenuate correlations. We believe an augmentation example, in which higher true correlations are more inflated than lower correlations, may be relevant in the AC context. In this view, assessors probably tend to notice that some pairs of dimensions are fairly highly related but that other pairs of dimensions are not (i.e., assessors develop implicit theories about dimension covariance). Use of these the-

[1] Marsh (1989) added external validity criteria to the MTMM models and correlated them with the method and trait factors. Although there was a strong a priori basis for assuming the external criteria to be more strongly related to the trait factors than to the method factors, the opposite result was found. In each of the matrices that were studied, the so-called method factors were more substantially correlated with an external validity criterion than were the trait factors. Therefore, Marsh concluded that the method factors seemed to represent to a certain extent trait variance instead of method variance and led to a miscalculation of dimension and method effects.

ories may then result in inflation of relationships between truly highly correlated dimensions (because these relationships are part of the implicit theory and therefore will be exaggerated) but will not result in inflation of truly unrelated dimensions.

Table A3 and Table B3 illustrate that, unlike the CFA combination model, the direct product model gives communalities for each variable. These communalities represent proportions of systematic variance in the measured variables accounted for by the model. In other words, there are no separate proportions of dimension and exercise variance, and, therefore, the relative size of dimension versus exercise effects cannot be addressed. This is because the direct product model, as a representation of Campbell and O'Connell's (1967, 1982) theoretical research, posits that the correlations between observed measures are not additive functions of trait and method effects but rather are multiplicative functions. Besides these communalities, the direct product model also gives estimates of dimension factor correlations and exercise factor correlations (see Table A3 and Table B3). These estimates are interpreted similarly as the factor correlations provided by the CFA combination model.

To our knowledge, the direct product model has not been applied to ACs. Some studies in the MTMM literature have found support (good fit and relative lack of estimation problems) for the direct product model as an alternative to the aforementioned CFA model (T. E. Becker & Vance, 1993; Coovert, Craiger, & Teachout, 1997; Goffin & Jackson, 1992). Other studies, however, have not found support for the direct product model (e.g., Bagozzi & Yi, 1990, 1991; T. E. Becker & Cote, 1994; Conway, 1996).

Correlated uniqueness model. This model includes additive dimension effects (represented by correlated dimension factors as in the CFA models described above) and exercise effects, but there are no exercise factors per se (Kenny, 1979; Kenny & Kashy, 1992; Marsh, 1989; Marsh & Bailey, 1991). Rather, to represent exercise variance, the unique factors (uniquenesses) of observed variables rated in the same exercise are allowed to be correlated. Precise estimates of proportions of exercise variance may then be obtained on the basis of the technique provided by Scullen (1999). Because there are no exercise factors, correlations among the exercise factors cannot be modeled, and this model therefore implicitly assumes uncorrelated exercises.

As illustrated with concrete examples in Table A4 and Table B4, the correlated uniqueness model gives estimates of dimension loadings, dimension factor correlations, and uniquenesses. These estimates are interpreted in a similar fashion as in the general CFA models. A particular feature of the correlated uniqueness model is that there are no exercise loadings. Instead, as we already noted, exercise variance estimates are derived from the correlations among the uniquenesses (from which this model gets its name). For example, in Table A4, the correlated uniquenesses (the off-diagonal elements in the uniqueness matrix) are very small—these values range from a low of .00 to a high of .15, which is consistent with the visual inspection of this MTMM matrix (see Table B1). Furthermore, when Scullen's (1999) technique is applied, the mean proportion of exercise variance is only .07. In combination with the mean proportion of dimension variance of .60, such results should be interpreted as support for the idea that AC ratings can be interpreted as measures of the intended dimensions and give no reason to question their validity for use in developmental feedback. In contrast, the results of Table B4 indicate that exercise variance (i.e., mean of squared exercise loadings = .51) dominates dimension variance (i.e., mean of squared dimension loadings = .26) and provide less support for the dimensions as building blocks of ACs.

In general, a large amount of recent evidence in the MTMM literature has shown that, as opposed to the aforementioned general CFA model, the correlated uniqueness model often produces good solutions (T. E. Becker & Cote, 1994; Conway, 1996; Kenny & Kashy, 1992; Marsh, 1989; Marsh & Bailey, 1991). Applications to ACs have generally replicated the good results (Kleinmann & Köller, 1997; Lievens & Van Keer, 1999; Sagie & Magnezy, 1997), although another AC study (Lance, Newbolt, et al., 2000) found estimation problems for the correlated uniqueness model.

Criteria for Model Appropriateness

Similar to previous studies (e.g., T. E. Becker & Cote, 1994; Conway, 1996), the appropriateness of the models was based on two sets of criteria. First, a model must not have produced an inadmissible solution such as a failure in the convergence of the iterative estimation procedure or parameter estimates held at boundary values to avoid improper estimates (e.g., negative variances, factor intercorrelations higher than unity).

As a second set of criteria for appropriateness, we used several goodness-of-fit indices. These indices included the relative noncentrality index (RNI), the Tucker–Lewis Index (TLI), and the root-mean-square error of approximation (RMSEA). The RNI and TLI measures of fit have been found to be unbiased and to be relatively independent of sample size (Marsh, Balla, & McDonald, 1988; McDonald & Marsh, 1990). The RMSEA is a measure of fit per degree of freedom of the model and is particularly attractive in that it allows the researcher to test the hypothesis of close fit (instead of perfect fit; Browne & Cudeck, 1993; Steiger, 1990). It is also possible to establish confidence intervals around the RMSEA. The criteria for evaluating these fit indices were for the TLI and the RNI to have values greater than or equal to .90 (T. E. Becker & Cote, 1994; Conway, 1996) and for the RMSEA to be less than or equal to .080.

Analyses

We used EQS Version 5.6 (Bentler, 1995) to analyze the CFA and correlated uniqueness models. MUTMUM (Browne, 1992) was used to conduct the direct product model analyses. Because direct product model analyses require a fully crossed design, the direct product model could be estimated for only 21 of the 34 matrices. Correlation, rather than covariance, matrices were analyzed because correlations were available for all studies. However, when possible, we analyzed both the covariance and correlation matrices and we compared the results. No notable differences were found (see also Cudeck, 1989). Maximum-likelihood estimation was used for all analyses. Automatic start values were used to fit each model, and if the model failed to converge after 500 iterations, the start values were set near estimates from other specifications that converged, and the analysis was repeated.

Results

Comparison of Different Models

The first question of central importance in this study was which structural equation model served as the best underlying representation of MTMM ratings of ACs. As we already mentioned, we started by testing several models that represented different conceptualizations of ACs. Table 2 displays the results of the various models along with their model appropriateness criteria. Table 2 shows that both simple models—namely, the dimensions-only model and the exercises-only model—showed a poor fit on average and that percentages of matrices with an acceptable fit for these two models were low (3% and 29%, respectively). The average fit statistics for the combination model with exercises and one general dimension just met the aforementioned fit criteria, providing an acceptable fit in 53% of the matrices. The remaining three combination models that specified both exercises and dimensions tended to fit much better. These results showed that, in terms of fit, AC ratings were best represented by a model with both exercises and dimensions.

Among the three combination models, the best fit values were obtained for the model with correlated dimensions and exercises specified as correlated uniqueness (i.e., a mean TLI of .987, a mean RNI of .985, and a mean RMSEA of .034). Inspection of the percentage of matrices yielding an acceptable fit revealed similar

Table 2
Summary of Model Performance for Various Models Across the 34 Multitrait–Multimethod Matrices

Model	Averages (N = 34)[a]					Summary		
	TLI	RNI	RMSEA	90% RMSEA confidence interval	No. of improper estimates	Matrices with acceptable fit[b]	Matrices with proper estimates	Matrices with proper estimates and acceptable fit
Simple models								
CFA—Correlated dimensions only	.452	.604	.216	.186–.244	4.18	3%	12%	3%
CFA—Correlated exercises only	.862	.889	.088	.055–.125	0.21	29%	85%	24%
Combination models								
CFA—Correlated exercises and one general dimension	.917	.945	.060	.031–.117	0.53	53%	59%	29%
CFA—Correlated exercises/correlated dimensions	.987	.981	.032	.016–.091	2.55	85%	9%	9%
Direct product model with correlated exercises/correlated dimensions	.978	.979	.043	.019–.099	0.76	81%	57%	52%
Correlated dimensions and exercises as correlated uniqueness	.987	.985	.034	.011–.086	1.00	88%	53%	53%

Note. TLI = Tucker–Lewis Index; RNI = relative noncentrality index; RMSEA = root-mean-square error of approximation; CFA = confirmatory factor analysis.
[a] These values represent mean values computed across the 34 matrices. Because the direct product model can be applied only for a fully crossed multitrait–multimethod design, the direct product model was estimated for only 21 matrices. [b] TLI and RNI $\geq$ .90 and RMSEA $\leq$ .08.

findings. The correlated uniqueness model scored best, with an acceptable fit in 88% of the matrices, slightly outperforming the CFA model with correlated dimensions and correlated exercises (85%) and the direct product model (81%).

Table 2 shows that there were larger differences between the combination models in terms of the criterion of producing solutions converging without improper parameter estimates (i.e., without estimates held at boundary values). In fact, the direct product and the correlated uniqueness models produced proper solutions in 57% and 53%, respectively, of the matrices. In contrast, in only 9% of the MTMM matrices the CFA model with correlated dimensions and correlated exercises produced a solution, which converged without improper parameter estimates. Further inspection of the parameter estimates of the CFA model indicated, for instance, the frequent occurrence of negative variances and negative dimension correlations. Some of the latter were even held at the boundary of −1.00. These negative dimension correlations are very counterintuitive and cast doubt on the results of this general CFA model (Coovert et al., 1997).

When both criteria were used (right-hand column of Table 2), the correlated uniqueness model was appropriate in 53% of the matrices and the direct product model in 52%. The CFA model with correlated dimensions and correlated exercises fared considerably worse (9%). The comparison between the correlated uniqueness model and the direct product model was not a particularly clean one because the direct product model results were based on only a subset of matrices. A more direct comparison involved only the 21 matrices for which the direct product model was estimated. For these matrices, the correlated uniqueness model actually performed better, with 15 appropriate solutions (71%) versus 11 (52%) for the direct product model.

In sum, given these results, we concluded that, in general, a model with correlated dimensions and with exercises specified as correlated uniquenesses was most appropriate for representing MTMM matrices of AC ratings. Hence, the next analyses used the estimates of this model to address the remaining two aims of our study.

Relative Importance of Exercise and Dimension Variance

Given the appropriateness of a combination model, our second aim was to estimate the relative importance of the dimensions versus the exercises across a large set of MTMM matrices of AC ratings. To this end, means for each type of estimate were first computed across all values within a matrix (e.g., all proportions of dimension variance estimates for a single matrix). As we already said, we hereby used the estimates of the best performing model (the correlated uniqueness model).

Table 3 presents the aggregate results as well as the individual results for each matrix. Across the 34 matrices, the mean proportion of variance obtained for dimensions was .34, and the mean proportion of variance for exercises was also .34. Thus, on average, dimensions and exercises explained equal proportions of variance. Yet, as indicated by the large variation in the mean proportion of dimension and exercise variance across the 34 matrices, this was not true for all ACs. To be more specific, the mean proportion of variance for dimensions varied from .17 to .62, and the mean proportion of exercise variance ranged from .07 to .69.

Table 3 also indicates that, on average, AC dimensions were substantially correlated. The average correlation among dimension factors was .71. Again, there was considerable variability among the MTMM matrices, with correlations ranging from .07 to .99.

These results are based on all MTMM matrices. Note, however, that our conclusions did not change much when only the 18 admissible correlated uniqueness solutions (those free of improper estimates) were considered. The proportion of dimension variance was somewhat higher (.38) than that for all matrices, but the conclusion that dimensions were virtually equally important as exercises still held. The similarity of this result attests to the fact that using all matrices did not invalidate our results.

Table 3
Mean Proportions of Dimension and Exercise Variance and Mean Dimension Factor Correlation for Correlated Uniqueness Model

Reference	Proportion of dimension variance	Proportion of exercise variance[a]	Dimension factor correlation
Arthur et al. (2000)	.60	.07	.55
A. S. Becker (1990)	.29	.41	.63
A. S. Becker (1990)	.27	.27	.78
Bobrow & Leonards (1997)	.25	.40	.87
Bycio et al. (1987)	.36	.40	.97
Chorvat (1994)	.40	.09	.33
Donahue et al. (1997)	.32	.32	.95
Fleenor (1996)	.29	.27	.80
Fredricks (1989)	.17	.56	.45
Harris et al. (1993)	.30	.23	.63
Joyce et al. (1994)	.24	.42	.58
Joyce et al. (1994)	.27	.24	.07
Kleinmann et al. (1994)	.62	.30	.77
Kleinmann et al. (1994)	.42	.35	.69
Kleinmann et al. (1996)	.37	.67	.56
Kleinmann (1997)	.44	.31	.90
Kleinmann (1997)	.54	.32	.92
Kleinmann (1997)	.36	.38	.82
Kolk et al. (2000)	.23	.41	.40
Kolk et al. (2000)	.23	.40	.73
Kudisch et al. (1997)	.35	.31	.43
Lievens & Van Keer (1999)	.36	.31	.73
Parker (1992)	.18	.23	.64
Robie et al. (2000)	.40	.25	.87
Sagie & Magnezy (1997)	.23	.31	.88
Sagie & Magnezy (1997)	.27	.28	.74
Schleicher et al. (1999)	.49	.30	.95
Schleicher et al. (1999)	.36	.39	.83
Schneider & Schmitt (1992)	.26	.51	.85
Sweeney (1976)	.31	.33	.95
Sweeney (1976)	.40	.36	.99
Van der Velde et al. (1994)	.17	.69	.77
Veldman (1994)	.34	.25	.54
Veldman (1994)	.50	.15	.59
M (*N* = 34)	.34	.34	.71
SD	.12	.13	.21
Range	.17–.62	.07–.69	.07–.99
95% confidence interval	.30–.38	.29–.38	.64–.78

[a] These means are based on proportion of exercise variance estimates obtained on the basis of the technique of Scullen (1999).

These results are more encouraging in terms of dimensions than are the results of earlier research (e.g., Bycio et al., 1987) that found exercise variance to predominate. However, equal proportions of dimension and exercise variance do not unambiguously support interpretation of AC ratings in terms of dimensions. In addition, the proportions of dimension variance also varied considerably in the ACs included in our evaluation. Therefore, it is important to try to identify design characteristics of ACs that account for the variability in proportions of dimension variance found. To this end, we now address our third question regarding what AC design characteristics are associated with higher proportions of dimension variance.

Impact of AC Design Characteristics

On the basis of theoretical and empirical grounds, we hypothesized significantly higher proportions of dimension variance for the following design characteristics: fewer numbers of dimensions, lower dimension–exercise ratio, use of behavioral checklists, transparent dimensions, psychologist assessors, longer assessor training program, and more similar exercises. To examine these hypotheses, we conducted *t* tests (see Table 4), using mean proportion of dimension variance for a matrix (averaged across all measures in the matrix) as the dependent variable. For ease of presentation, median splits were done for the number of dimensions (*Mdn* = 5.5) and the ratio of dimensions to exercises (*Mdn* = 1.5). We also conducted exploratory *t* tests using mean exercise variance as the dependent variable. Because the correlated uniqueness model produced much better solutions and presumably more reliable parameter estimates than did the other models, we used proportions of dimension variance and exercise variance estimates only from the correlated uniqueness model. Note again that correlated uniqueness model results of all 34 MTMM matrices were included. This was done to increase the power of the *t* tests. When only the 18 appropriate correlated uniqueness solutions

Table 4
Results for t Tests for Differences in Proportions of Dimension and Exercise Variance From Correlated Uniqueness Model

Variable	No. of studies	Proportion of dimension variance	Proportion of exercise variance[a]
No. of dimensions			
5 or fewer	20	.37*	.35
6 or more	14	.30*	.33
Dimension–exercise ratio			
Low	14	.38	.34
High	20	.31	.33
Behavioral checklists			
Used	16	.37	.35
Not used	15	.31	.36
Transparent dimensions			
Transparent	4	.35	.35
Not transparent	29	.35	.34
Type of assessor			
Managers	14	.27*	.38
Psychologists	16	.39*	.34
Training length			
More than 1 day	17	.30*	.33
1 day or less	10	.45*	.34
Exercise similarity			
Low	23	.33	.34
High	11	.36	.34

[a] These means are based on proportion of exercise variance estimates obtained on the basis of the technique of Scullen (1999).
* $p < .05$.

were considered, results were similar to those presented below, although p values for the t tests became larger because of the smaller n.

Table 4 shows support for some of the hypotheses. Significantly ($p < .05$) higher proportions of dimension variance were found when fewer dimensions were used and when psychologists served as assessors. Higher dimension variance proportions were also noted for more similar exercises, when the dimension–exercise ratio was low, and when behavioral checklists were used. Yet, the differences were not significant. Use of transparent dimensions yielded virtually no effects. Contrary to our hypothesis, longer training was associated with significantly smaller proportions of dimension variance.

To address possible confounding of the AC design characteristics, we computed correlations between all dichotomous design variables (to save space, the correlation matrix is not presented). There were substantial correlations (>.40) among all three variables showing differences in proportion of dimension variance: number of dimensions, training, and type of assessor. We therefore conducted a multiple regression analysis with these variables as predictors and proportion of dimension variance as the criterion ($N = 34$, with matrix as the unit of analysis). To increase power, we used the continuous number of ratings variable (not the dichotomized variable that we used for the t test). In this analysis, each predictor variable had at least a marginally significant ($p < .10$) regression coefficient. Given the small N, this finding provides good evidence that our significant t-test results are not spurious results of confounding with our other design characteristics.

Discussion

This study's large-scale and systematic evaluation of MTMM matrices of AC ratings yields several important conclusions with regard to the underlying model for ACs, the relative sizes of dimension and exercise effects, AC design characteristics, and, most important, the interpretation and use of AC ratings.

Dimensions and Exercises as the Best Conceptualization of ACs

A first series of results contributes to the issue of the appropriate conceptualization of ACs. Among the models, which represented different conceptualizations of ACs, excellent fit results were found for combination models, which included both dimensions and exercises. It is particularly important that the fit of these models was better than that for the simple dimensions-only and exercises-only models or the model with exercises and one global dimension. This result supports the use of dimensions as essential building blocks of ACs and refutes the argument that ACs constitute nothing more than a series of work samples (Robertson et al., 1987).

A related important finding is that, on average, exercises and dimensions explained equal amounts of variance in AC ratings. Yet, the large variability associated with dimension variance proportions across the ACs considered also shows that ACs varied widely in terms of the extent to which they were dimension-based. In other words, some ACs were better than others in measuring dimensions. The implication of this finding is that each AC needs to be validated to interpret its results in terms of dimension versus exercise variance. If such a validation shows that the dimensions (in addition to the AC exercises) emerge as important factors, then it is warranted to build assessees' reports and feedback around the dimensions.

The fact that the exercises emerged as important factors and that they were, on average, as important as dimensions can be inter-

preted in two different ways. On the one hand, considerable proportions of exercise variance may be regarded as sources of measurement method bias. This is the traditional explanation of method effects in ACs, which detracts from their construct validity. According to this view, it is disappointing that we were not able to identify AC design characteristics, which decrease these unwanted method effects. On the other hand, proportions of method variance may also be considered as sources of valid variance. In fact, Lance, Newbolt, et al. (2000) recently correlated exercise factors with external correlates such as cognitive ability and job performance measures. They found support for the view that some AC exercise factors represent valid cross-situational specificity in AC performance instead of method bias. On the basis of this interpretation, it is good news that exercise variance remains virtually unchanged under the various AC design characteristics. However, the relatively strong exercise effects have implications for use of dimensions to provide feedback: Even when feedback is organized around dimensions, the dimensions should not be the sole focus. Rather, dimension feedback should be somewhat context-specific. For example, an assessee may generally need to improve leadership skills but may have a need to concentrate especially on group situations as opposed to one-on-one situations.

According to the best performing model (the correlated uniqueness model), the exercise effects are best specified in MTMM analyses as correlated uniquenesses instead of as separate exercise factors. Implicit in specifying exercise effects as correlated uniquenesses is that only correlations are allowed among the uniquenesses of measures of the same method (Lance, Noble, & Scullen, 2000) and, therefore, that AC exercises are considered to be independent methods. The question is whether this assumption is realistic for ACs. The answer to this question mainly depends on the AC under investigation. For example, this assumption may be unrealistic for ACs that consist of similar exercises such as two group discussions. In contrast, it may be more tenable for ACs that are composed of dissimilar exercises such as an in-basket and a group discussion. In general, we believe that AC exercises are relatively independent from each other because AC designers tend to choose simulation exercises that sample different job situations. In addition, different raters typically assess candidates in various simulation exercises. Research also shows that assessors perceive AC exercises to be quite different situations (Highhouse & Harris, 1993).

AC Design Characteristics

A second series of results contributes to research on the impact of AC design considerations on the quality of dimension measurement in ACs and helps to explain why dimensions are better measured in some ACs than in others. In particular, our large-scale evaluation demonstrated that dimension variance can be significantly increased when fewer dimensions are used and when psychologists serve as assessors. Moreover, in all of these design conditions, dimension variance slightly dominated exercise variance (see Table 4). From a theoretical point of view, these results provide support for the limited cognitive capacity perspective and the expert assessor perspective discussed in the introduction. For practitioners, these findings are important because they indicate that careful AC design can help to substantially increase the quality of the AC in general and the confidence in dimension-based inferences in particular. Along these lines, it should be noted that each of our analyses looked at a single design characteristic. Thus, if an AC user or designer implements multiple characteristics, the effect on dimension variance is likely to be even larger.

The most straightforward explanation for the positive effects of using a limited set of dimensions (and a low dimension–exercise ratio) is that reducing the dimensions to be rated per exercise makes the task of assessors less cognitively complex. Another explanation is that if fewer dimensions are used, there is less chance that different dimensions are rated on the basis of the same or linked behaviors (Kleinmann et al., 1995). Or to put it differently, use of a limited set of dimensions will then inherently lead to the use of more specific and conceptually distinct dimensions, increasing the variance due to dimensions. However, this is not necessarily true because a small number of dimensions may also lead to the use of broad dimensions (i.e., dimensions at a higher level of abstraction). The latter case is a possible pitfall that practitioners should avoid at all times. After all, when more broad dimensions are used, behaviors tend to overlap considerably across those dimensions so that assessors are overworked when distinguishing among them. Taken together, when implementing the design consideration of a limited set of dimensions, practitioners should select those dimensions that are most relevant for the target job. This limited set of dimensions should also be sufficiently specific and conceptually distinct from each other so that the use of broad dimensions is avoided.

Our finding of psychologist assessors providing a better measurement of AC dimensions than managerial assessors is also in line with prior AC research (Lievens, in press; Sagie & Magnezy, 1997). Psychologists' educational background and job practices, which make them increasingly familiar with individual-differences constructs, may explain this superiority. Given these consistent findings, we recommend that psychologists play a key role in assessor teams of ACs. For example, psychologists may serve as coaches of managerial assessors or as chairs of the discussion session. This recommendation is especially relevant for ACs conducted for developmental purposes, which require the use of dimension scores for feedback purposes.

Another finding is that the use of more similar exercises and behavioral checklists may also lead to higher dimension variance proportions, although the increases do not reach statistical significance. Therefore, the impact of these design considerations should be investigated further before firm conclusions are drawn. With respect to behavioral checklists, future studies could, for instance, investigate whether the number of behaviors listed, the grouping of these behaviors, and the inclusion of a retranslation procedure designed to eliminate "fuzzy" behaviors make a difference in terms of valid dimension measurement.

A counterintuitive finding is that shorter assessor training programs (<1 day) result in higher dimension variance, whereas longer training programs (>1 day) are related to lower dimension variance. A possible explanation is that the length of assessor training is not really important. However, given the emphasis on training in AC practice, this explanation does not seem very likely. Perhaps the quality of dimension measurement is more affected by the type of training than by the length of such training. Recent studies (Lievens, 2001; Schleicher et al., 1999) have supported this explanation, revealing that a training program that imposed a common frame-of-reference on assessors was most promising. Another possible explanation is that the length of training reported

in the studies is misleading. In fact, the length of assessor training depends not only on the thoroughness of teaching assessors crucial observation and evaluation skills but also on the number of AC exercises—with more exercises leading to longer training programs (Task Force on Assessment Center Guidelines, 1989).

Limitations

In this study, the correlated uniqueness model represented best MTMM matrices of AC ratings. A limitation of this correlated uniqueness model is that if the restriction of independent methods does not hold, it suffers from a biasing effect (i.e., inflation of trait variances; Kenny & Kashy, 1992; Lance, Noble, & Scullen, 2000). Simulation studies of Marsh and Bailey (1991, p. 65), however, found no real evidence that this upward bias was substantial, because correlated uniqueness parameter estimates were accurate even when method correlations were as high as .60. In addition, the present study's results provide a test of whether dimension variance is inflated by use of the correlated uniqueness model and its assumption of independent methods because we included similarity of AC exercises as a design variable. More precisely, when exercises were not similar (more independent methods), dimension variance was .33 (see Table 4). When exercises were similar (less independent methods), the mean proportion of dimension variance was .36. Whereas the insignificant rise in dimension variance for similar exercises (.36) may also be explained by upward bias due to the correlated uniqueness model, it is not likely that the value of .33 found for dissimilar exercises was inflated. Because the latter value closely parallels the general correlated uniqueness estimate of dimension variance (.34; see Table 3), all evidence attests to the accuracy of the correlated uniqueness estimates. Yet, future research should use Monte Carlo simulations to examine more thoroughly the possible upward bias of the correlated uniqueness model.

Directions for Future Research

Future research endeavors in this domain should be geared toward the following directions. First, because of the high variability in dimension variance across the ACs, it is pivotal that practitioners and researchers find out what other aspects of AC design (besides this study's design characteristics) lead to more dimension-based ratings. These design considerations can help to further explain why some ACs are better able to measure dimensions and therefore have more dimension variance than others. It is also important that these design considerations are grounded by theory. Our results show that design changes inspired by the limited cognitive capacity perspective and the expert assessor perspective hold the most promise.

Second, future studies should use a broad validation design to determine the effects of AC design characteristics. This study's design recommendations were formulated solely in terms of increasing the quality of construct measurement of ACs. Future studies are needed to ascertain whether the suggested AC procedural interventions also positively affect content and criterion-related validity. A particularly intriguing question is how proportions of dimension versus exercise variance relate to criterion-related validity. For instance, do ACs with high exercise variance and low dimension variance show more criterion-related validity than ACs with low exercise variance and high dimension variance? Or is the highest criterion-related validity obtained for ACs that show both high exercise variance and high dimension variance? The latter result would be most consistent with the original foundations of ACs.

Finally, studies are needed that decompose variance according to the three main sources of variance in AC ratings: dimensions, exercises, and assessors. A limitation inherent in all models currently tested (i.e., the CFA, correlated uniqueness, and direct product models) is that they recognize only two sources of systematic variance, namely, dimensions and exercises. Exercise variance is then necessarily a combination of variance due to different exercises and variance due to different assessors. To disentangle dimension, exercise, and assessor variance, different assessors have to provide dimensional ratings of the same assessees in each exercise. If such a design were available, hierarchical CFA (Lance, Teachout, & Donnelly, 1992; Marsh & Hocevar, 1988) could be applied. This model would provide even more detailed construct validity information than that provided in the present study.

Conclusion

In the past, there has been considerable debate about the poor construct (i.e., dimension) measurement in ACs. Our large-scale and systematic evaluation of existing MTMM matrices of AC ratings shows that individual-differences constructs as operationalized by AC dimensions do have their place in the AC framework. However, we also found that the quality of dimension measurement depends on the AC under investigation. Dimensions are significantly better measured when psychologists serve as assessors and when the inclusion of fewer dimensions makes the assessor task less complex. Use of behavioral checklists, a lower dimension–exercise ratio, and similar exercises also increases dimension variance.

References

References marked with an asterisk indicate studies providing multitrait–multimethod matrices of assessment center ratings.

*Arthur, W. A., Jr., Woehr, D. J., & Maldegen, R. (2000). Convergent and discriminant validity of assessment center dimensions: A conceptual and empirical re-examination of the assessment center construct-related validity paradox. *Journal of Management, 26*, 813–835.

Bagozzi, R. P., & Yi, Y. (1990). Assessing method variance in multitrait–multimethod matrices: The case of self-reported affect and perceptions at work. *Journal of Applied Psychology, 75*, 547–560.

Bagozzi, R. P., & Yi, Y. (1991). Multitrait–multimethod matrices in consumer research. *Journal of Consumer Psychology, 17*, 426–439.

*Becker, A. S. (1990). *The effects of a reduction in assessor roles on the convergent and discriminant validity of assessment center ratings.* Unpublished doctoral dissertation, University of Missouri—St. Louis.

Becker, T. E., & Cote, J. A. (1994). Additive and multiplicative method effects in applied psychological research: An empirical assessment of three methods. *Journal of Management, 20*, 625–641.

Becker, T. E., & Vance, R. J. (1993). Construct validity of three types of organizational citizenship behavior: An illustration of the direct product model with refinements. *Journal of Management, 19*, 663–682.

Bentler, P. M. (1995). *EQS: Structural equations program manual.* Encino, CA: Multivariate Software.

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology, 74*, 478–494.

*Bobrow, W., & Leonards, J. S. (1997). Development and validation of an

assessment center during organizational change. *Journal of Social Behavior and Personality, 12*, 217–236.

Brannick, M. T., & Spector, P. E. (1990). Estimation problems in the block-diagonal model of the multitrait–multimethod matrix. *Applied Psychological Measurement, 14*, 325–339.

Browne, M. W. (1984). The decomposition of multitrait–multimethod matrices by generalized least squares. *British Journal of Mathematical and Statistical Psychology, 37*, 1–21.

Browne, M. W. (1992). *MUTMUM user's guide*. Columbus: Ohio State University, Department of Psychology.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

*Bycio, P., Alvares, K. M., & Hahn, J. (1987). Situational specificity in assessment center ratings: A confirmatory factor analysis. *Journal of Applied Psychology, 72*, 463–474.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin, 56*, 81–105.

Campbell, D. T., & O'Connell, E. J. (1967). Method factors in multitrait–multimethod matrices: Multiplicative rather than additive? *Multivariate Behavioral Research, 2*, 409–426.

Campbell, D. T., & O'Connell, E. J. (1982). Methods as diluting trait relationships rather than adding irrelevant systematic variance. In D. Brinberg & L. H. Kidder (Eds.), *New directions for methodology of social and behavioral science: Forms of validity in research* (pp. 93–111). San Francisco: Jossey-Bass.

Cardy, R. L., Bernardin, H. J., Abbott, J. G., Senderak, M. P., & Taylor, K. (1987). The effects of individual performance schemata and dimension familiarization on rating accuracy. *Journal of Occupational and Organizational Psychology, 60*, 197–205.

Chi, M. T. H., Glaser, R., & Farr, M. J. (1988). *The nature of expertise*. Hillsdale, NJ: Erlbaum.

*Chorvat, V. P. (1994). *Toward the construct validity of assessment center leadership dimensions: A multitrait–multimethod investigation using confirmatory factor analysis*. Unpublished doctoral dissertation, University of South Florida.

Conway, J. M. (1996). Analysis and design of multitrait–multirater performance appraisal studies. *Journal of Management, 22*, 139–162.

Coovert, M. D., Craiger, J. P., & Teachout, M. S. (1997). Effectiveness of the direct product versus confirmatory factor model for reflecting the structure of multimethod–multirater job performance data. *Journal of Applied Psychology, 82*, 271–280.

Cudeck, R. (1989). Analysis of correlation matrices using covariance structure models. *Psychological Bulletin, 105*, 317–327.

*Donahue, L. M., Truxillo, D. M., Cornwell, J. M., & Gerrity, M. J. (1997). Assessment center construct validity and behavioral checklists: Some additional findings. *Journal of Social Behavior and Personality, 12*, 85–108.

Dugan, B. (1988). Effects of assessor training on information use. *Journal of Applied Psychology, 73*, 743–748.

Fiske, S. T., & Taylor, S. E. (1991). *Social cognition*. Singapore: McGraw-Hill.

*Fleenor, J. W. (1996). Constructs and developmental assessment centers: Further troubling empirical findings. *Journal of Business and Psychology, 10*, 319–333.

*Fredricks, A. J. (1989). *Assessment center ratings: Models and process*. Unpublished doctoral dissertation, University of Nebraska—Lincoln.

Gaugler, B. B., & Thornton, G. C. (1989). Number of assessment center dimensions as a determinant of assessor accuracy. *Journal of Applied Psychology, 74*, 611–618.

Goffin, R. D., & Jackson, D. N. (1992). Analysis of multitrait–multirater performance appraisal data: Composite direct product method versus confirmatory factor analysis. *Multivariate Behavioral Research, 27*, 363–385.

*Harris, M. M., Becker, A. S., & Smith, D. E. (1993). Does the assessment center scoring method affect the cross-situational consistency of ratings? *Journal of Applied Psychology, 78*, 675–678.

Highhouse, S., & Harris, M. M. (1993). The measurement of assessment center situations: Bem's template matching technique for examining exercise similarity. *Journal of Applied Social Psychology, 23*, 140–155.

*Joyce, L. W., Thayer, P. W., & Pond, S. B. (1994). Managerial functions: An alternative to traditional assessment center dimensions? *Personnel Psychology, 47*, 109–121.

Kenny, D. A. (1979). *Correlation and causality*. New York: Wiley.

Kenny, D. A., & Kashy, D. A. (1992). Analysis of the multitrait–multimethod matrix by confirmatory factor analysis. *Psychological Bulletin, 112*, 165–172.

*Kleinmann, M. (1997). Transparenz der Anforderungsdimensionen: Ein Moderator der Konstrukt- und Kriteriumsvalidität des Assessment-Centers [Transparency of the required dimensions: A moderator of assessment centers' construct and criterion validity]. *Zeitschrift für Arbeits- und Organisationspsychologie, 41*, 171–181.

*Kleinmann, M., Andres, J., Fedtke, C., Godbersen, F., & Köller, O. (1994). Der Einfluss unterschiedlicher auswertungsmethoden auf die Konstruktvalidität von Assessment-Centern [The influence of different rating procedures on the construct validity of assessment center methods]. *Zeitschrift für Experimentelle und Angewandte Psychologie, 41*, 184–210.

Kleinmann, M., Exler, C., Kuptsch, C., & Köller, O. (1995). Unabhangigkeit und Beobachtbarkeit von Anforderungsdimensionen im Assessment Center als Moderatoren der Konstruktvaliditat [Independence and observability of dimensions as moderators of construct validity in the assessment center]. *Zeitschrift für Arbeits- und Organisationspsychologie, 39*, 22–28.

*Kleinmann, M., & Köller, O. (1997). Construct validity of assessment centers: Appropriate use of confirmatory factor analysis and suitable construction principles. *Journal of Social Behavior and Personality, 12*, 65–84.

*Kleinmann, M., Kuptsch, C., & Köller, O. (1996). Transparency: A necessary requirement for the construct validity of assessment centres. *Applied Psychology: An International Review, 45*, 67–84.

Klimoski, R. J. (1993). Predictor constructs and their measurement. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 99–134). San Francisco: Jossey-Bass.

*Kolk, N. J., Born, M. P., & Van der Flier, H. (2000, April). *The transparent assessment center: The effect of revealing dimensions to applicants*. Paper presented at the 15th Annual Conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

*Kudisch, J. D., Ladd, R. T., & Dobbins, G. H. (1997). New evidence on the construct validity of diagnostic assessment centers: The findings may not be so troubling after all. *Journal of Social Behavior and Personality, 12*, 129–144.

Lance, C. E., Newbolt, W. H., Gatewood, R. D., Foster, M. S., French, N. R., & Smith, D. E. (2000). Assessment center exercises represent cross-situational specificity, not method bias. *Human Performance, 13*, 323–353.

Lance, C. E., Noble, C. L., & Scullen, S. E. (2000, April). *The merits of the correlated uniqueness model for multitrait–multimethod data have been oversold*. Paper presented at the 15th Annual Conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Lance, C. E., Teachout, M. S., & Donnelly, T. M. (1992). Specification of the criterion construct space: An application of hierarchical confirmatory factor analysis. *Journal of Applied Psychology, 77*, 437–452.

Landy, F. J., Shankster, L. J., & Kohler, S. S. (1994). Personnel selection and placement. *Annual Review of Psychology, 46*, 261–296.

Levin, J. (1965). Three-mode factor analysis. *Psychological Bulletin, 64*, 442–452.

Lievens, F. (1998). Factors which improve the construct validity of assess-

ment centers: A review. *International Journal of Selection and Assessment, 6,* 141–152.

Lievens, F. (2001). Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. *Journal of Applied Psychology, 86,* 255–264.

Lievens, F. (in press). Assessors and use of assessment center dimensions: A fresh look at a troubling issue. *Journal of Organizational Behavior.*

*Lievens, F., & Van Keer, E. (1999, May). *Modeling method effects in assessment centers: An application of the correlated uniqueness approach.* Paper presented at the 14th Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Lorenzo, R. V. (1984). Effects of assessorship on managers' proficiency in acquiring, evaluating, and communicating information about people. *Personnel Psychology, 37,* 617–634.

Lowry, P. E. (1997). The assessment center process: New directions. *Journal of Social Behavior and Personality, 12,* 53–62.

Magnusson, D. (1982). *Toward a psychology of situations: An interactional perspective.* Hillsdale, NJ: Erlbaum.

Maher, P. T. (1990, March). *How many dimensions are enough?* Paper presented at the International Congress on the Assessment Center Method, Orange, CA.

Maher, P. T. (1995, May). *An analysis of the impact of the length of assessor training on assessor competency.* Paper presented at the International Congress on the Assessment Center Method, Kansas City, KS.

Marsh, H. W. (1989). Confirmatory factor analyses of multitrait–multimethod data: Many problems and a few solutions. *Applied Psychological Measurement, 13,* 335–361.

Marsh, H. W., & Bailey, M. (1991). Confirmatory factor analysis of multitrait–multimethod data: A comparison of alternative models. *Applied Psychological Measurement, 15,* 47–70.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103,* 391–410.

Marsh, H. W., & Hocevar, D. (1988). A new procedure for analysis of multitrait–multimethod data: An application of second order confirmatory factor analysis. *Journal of Applied Psychology, 73,* 107–111.

McDonald, R. P., & Marsh, H. W. (1990). Choosing a multivariate model: Noncentrality and goodness-of-fit. *Psychological Bulletin, 107,* 247–255.

Neidig, R. D., & Neidig, P. J. (1984). Multiple assessment center exercises and job relatedness. *Journal of Applied Psychology, 69,* 182–186.

*Parker, M. W. (1992). *A construct validation of the Florida Principal Competencies Assessment Center using confirmatory factor analysis.* Unpublished doctoral dissertation, University of South Florida.

Reilly, R. R., Henry, S., & Smither, J. W. (1990). An examination of the effects of using behavior checklists on the construct validity of assessment center dimensions. *Personnel Psychology, 43,* 71–84.

Robertson, I., Gratton, L., & Sharpley, D. (1987). The psychometric properties and design of managerial assessment centres: Dimensions into exercises won't go. *Journal of Occupational Psychology, 60,* 187–195.

*Robie, C., Adams, K. A., Osburn, H. G., Morris, M. A., & Etchegaray, J. M. (2000). Effects of the rating process on the construct validity of assessment center dimension evaluations. *Human Performance, 13,* 355–370.

Sackett, P. R., & Dreher, G. F. (1982). Constructs and assessment center dimensions: Some troubling empirical findings. *Journal of Applied Psychology, 67,* 401–410.

*Sagie, A., & Magnezy, R. (1997). Assessor type, number of distinguishable dimension categories, and assessment centre construct validity. *Journal of Occupational and Organizational Psychology, 70,* 103–108.

*Schleicher, D. J., Day, D. V., Mayes, B. T., & Riggio, R. E. (1999, May). *A new frame for frame-of-reference training: Enhancing the construct validity of assessment centers.* Paper presented at the 14th Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Schmitt, N., & Chan, D. (1998). *Personnel selection: A theoretical approach.* Thousand Oaks, CA: Sage.

Schmitt, N., & Stults, D. M. (1986). Methodology review: Analysis of multitrait–multimethod matrices. *Applied Psychological Measurement, 10,* 1–22.

*Schneider, J. R., & Schmitt, N. (1992). An exercise design approach to understanding assessment center dimension and exercise constructs. *Journal of Applied Psychology, 77,* 32–41.

Scullen, S. E. (1999). Using confirmatory factor analysis of correlated uniquenesses to estimate method variance in multitrait–multimethod matrices. *Organizational Research Methods, 2,* 275–292.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25,* 173–180.

*Sweeney, D. C. (1976). *The development and analysis of rating scales for the Chicago Police Recruit Assessment Center.* Unpublished manuscript, Bowling Green State University.

Tanaka, J. S. (1987). "How big is enough?": Sample size and goodness of fit in structural equation models with latent variables. *Child Development, 58,* 134–146.

Task Force on Assessment Center Guidelines. (1989). Guidelines and ethical considerations for assessment center operations. *Public Personnel Management, 18,* 457–470.

Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika, 31,* 279–311.

*Van der Velde, E. G., Born, M. P., & Hofkes, K. (1994). Begripsvalidering van een assessment center met behulp van confirmatorische factoranalyse [Construct validity of an assessment center using confirmatory factor analysis]. *Gedrag en Organisatie, 7,* 18–26.

*Veldman, W. M. (1994). *Assessment centers and candidates' personal qualities: A study on the correlations between assessment center ratings.* Unpublished doctoral dissertation, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands.

Viswesvaran, C., & Ones, D. S. (1995). Theory testing: Combining psychometric meta-analysis and structural equations modeling. *Personnel Psychology, 48,* 865–885.

Widaman, F. (1985). Hierarchically nested covariance structure models for multitrait–multimethod data. *Applied Psychological Measurement, 9,* 1–26.

Wothke, W., & Browne, M. W. (1990). The direct product model for the MTMM matrix parameterized as a second order factor analysis model. *Psychometrika, 55,* 255–262.

Zedeck, S. (1986). A process analysis of the assessment center method. *Research in Organizational Behavior, 8,* 259–296.

Appendix A

Example of Multitrait–Multimethod and Structural Equation Analyses of a Matrix in Which Dimensions Emerge as Dominant Factors

Table A1
Multitrait–Multimethod Matrix of Arthur et al. (2000)

	IB				AL				MP			
Variable	OC	TB	IN	ST	OC	TB	IN	ST	OC	TB	IN	ST
IB												
OC	—											
TB	.43	—										
IN	.38	.48	—									
ST	.31	.22	.30	—								
AL												
OC	**.63**	.40	.39	.27	—							
TB	.37	**.55**	.42	.22	.51	—						
IN	.30	.40	**.53**	.17	.45	.51	—					
ST	.32	.20	.24	**.72**	.38	.28	.24	—				
MP												
OC	**.54**	.38	.42	.24	**.68**	.48	.45	.31	—			
TB	.36	**.44**	.44	.22	.49	**.62**	.41	.28	.60	—		
IN	.31	.26	**.50**	.23	.39	.41	**.49**	.27	.49	.51	—	
ST	.34	.21	.28	**.70**	.34	.28	.22	**.80**	.38	.32	.30	—

Note. The mean heterotrait–monomethod correlation is .39, the mean monotrait–heteromethod correlation (individual values are in boldface) is .60, and the mean heterotrait–heteromethod correlation is .33. IB = in-basket exercise; AL = allocation exercise; MP = management problems exercise; OC = oral communication; TB = team building; IN = innovation; ST = stress tolerance. Adapted from *Journal of Management, 26,* W. A. Arthur, Jr., D. J. Woehr, & R. Maldegen, "Convergent and Discriminant Validity of Assessment Center Dimensions: A Conceptual and Empirical Re-examination of the Assessment Center Construct-Related Validity Paradox," pp. 813–835, Copyright 2000, with permission from Elsevier Science.

(Appendixes continue)

Table A2

General Confirmatory Factor Analysis Model Parameter Estimates for Combination Model Applied to Multitrait–Multimethod Matrix of Arthur et al. (2000)

	Dimension loadings				Exercise loadings			
Variable	OC	TB	IN	ST	IB	AL	MP	Uniqueness
IB								
OC	.74				.25			.39
TB		.76			.25			.37
IN			.64		.55			.27
ST				.77	.26			.33
AL								
OC	.77					.44		.22
TB		.67				.51		.31
IN			.67			.40		.39
ST				.88		.28		.15
MP								
OC	.59						.60	.28
TB		.47					.70	.29
IN			.45				.56	.49
ST				.83			.32	.21

	Factor correlations						
Factor	OC	TB	IN	ST	IB	AL	MP
OC	—						
TB	.62	—					
IN	.54	.66	—				
ST	.38	.27	.27	—			
IB					—		
AL					.49	—	
MP					.64	.84	—

Note. $\chi^2(33, N = 149) = 13.51$, Tucker–Lewis Index = 1.050, relative noncentrality index = 1.00, root-mean-square error of approximation = 0.00. OC = oral communication; TB = team building; IN = innovation; ST = stress tolerance; IB = in-basket exercise; AL = allocation exercise; MP = management problems exercise.

Table A3
Direct Product Model Parameter Estimates for Combination Model Applied to Multitrait–Multimethod Matrix of Arthur et al. (2000)

	Communalities		
Variable	IB	AL	MP
OC	.84	.88	.88
TB	.79	.84	.84
IN	.74	.79	.80
ST	.93	.95	.95

	Dimension correlations			
	OC	TB	IN	ST
OC	—			
TB	.72	—		
IN	.66	.77	—	
ST	.42	.34	.36	—

	Exercise correlations		
	IB	AL	MP
IB	—		
AL	.83	—	
MP	.76	.87	—

Note. $\chi^2(51, N = 149) = 29.18$, Tucker–Lewis Index = 1.032, relative noncentrality index = 1.00, root-mean-square error of approximation = 0.00. IB = in-basket exercise; AL = allocation exercise; MP = management problems exercise; OC = oral communication; TB = team building; IN = innovation; ST = stress tolerance.

(*Appendixes continue*)

Table A4

Correlated Uniqueness Model Parameter Estimates for Combination Model Applied to Multitrait–Multimethod Matrix of Arthur et al. (2000)

					Uniquenesses											
	Dimension loadings				IB				AL				MP			
Variable	OC	TB	IN	ST	OC	TB	IN	ST	OC	TB	IN	ST	OC	TB	IN	ST
IB																
OC	.71				.51											
TB		.65			.12	.59										
IN			.73		.07	.14	.48									
ST				.79	.05	.06	.09	.37								
AL																
OC	.87								.24							
TB		.83							.00	.31						
IN			.73						.02	.06	.47					
ST				.91					.04	.01	.03	.18				
MP																
OC	.77												.39			
TB		.73											.15	.45		
IN			.66										.10	.13	.54	
ST				.88									.07	.04	.03	.22

	Factor correlations			
Factor	OC	TB	IN	ST
OC	—			
TB	.72	—		
IN	.67	.74	—	
ST	.44	.37	.38	—

Note. $\chi^2(30, N = 149) = 16.08$, Tucker–Lewis Index = 1.040, relative noncentrality index = 1.00, root-mean-square error of approximation = 0.00. IB = in-basket exercise; AL = allocation exercise; MP = management problems exercise; OC = oral communication; TB = team building; IN = innovation; ST = stress tolerance.

Appendix B

Example of Multitrait–Multimethod and Structural Equation Analyses of a Matrix in Which Exercises Emerge as Dominant Factors

Table B1
Multitrait–Multimethod Matrix of Schneider and Schmitt (1992)

	GR1			GR2			RP1			RP2		
Variable	PS	IP	IN	PS	IP	IN	PS	IP	IN	PS	IP	IN
GR1												
PS	—											
IP	.66	—										
IN	.81	.60	—									
GR2												
PS	**.15**	.11	.19	—								
IP	.26	**.22**	.28	.60	—							
IN	.26	.19	**.33**	.76	.69	—						
RP1												
PS	**.27**	.11	.28	**.32**	.27	.39	—					
IP	.25	**.07**	.28	.06	**.20**	.13	.70	—				
IN	.13	−.06	**.16**	.14	.13	**.25**	.77	.65	—			
RP2												
PS	**.21**	.17	.22	**.27**	.26	.19	**.34**	.35	.25	—		
IP	.24	**.17**	.29	.17	**.31**	.18	.33	**.49**	.30	.78	—	
IN	.25	.19	**.25**	.16	.23	**.14**	.35	.44	**.32**	.84	.79	—

Note. The mean heterotrait–monomethod correlation is .72, the mean monotrait–heteromethod correlation (individual values are in boldface) is .25, and the mean heterotrait–heteromethod correlation is .22. GR1 = first group discussion exercise; GR2 = second group discussion exercise; RP1 = first role-playing exercise; RP2 = second role-playing exercise; PS = problem solving; IP = interpersonal skills; IN = initiative. From "An Exercise Design Approach to Understanding Assessment Center Dimension and Exercise Constructs," by J. R. Schneider and N. Schmitt, 1992, *Journal of Applied Psychology, 77,* p. 37. Copyright 1992 by the American Psychological Association. Adapted with permission.

(*Appendixes continue*)

Table B2

General Confirmatory Factor Analysis Model Parameter Estimates for Combination Model Applied to Multitrait–Multimethod Matrix of Schneider and Schmitt (1992)

	Dimension loadings			Exercise loadings				
Variable	PS	IP	IN	GR1	GR2	RP1	RP2	Uniqueness
GR1								
PS	.22			.89				.14
IP		.02		.71				.49
IN			.26	.83				.22
GR2								
PS	.41				.71			.32
IP		.44			.64			.39
IN			.54		.78			.10
RP1								
PS	.99					.13		.01
IP		.86				.32		.15
IN			.82			.20		.29
RP2								
PS	.29						.88	.15
IP		.38					.79	.22
IN			.32				.88	.15

	Factor correlations						
Factor	PS	IP	IN	GR1	GR2	RP1	RP2
PS	—						
IP	.77	—					
IN	.92	.83	—				
GR1				—			
GR2				.20	—		
RP1				.04	−1.00	—	
RP2				.21	.09	.59	—

Note. $\chi^2(33, N = 89) = 27.10$, Tucker–Lewis Index = 1.018, relative noncentrality index = 1.00, root-mean-square error of approximation = 0.00. PS = problem solving; IP = interpersonal skills; IN = initiative; GR1 = first group discussion exercise; GR2 = second group discussion exercise; RP1 = first role-playing exercise; RP2 = second role-playing exercise.

Table B3
Direct Product Model Parameter Estimates for Combination Model Applied to Multitrait–Multimethod Matrix of Schneider and Schmitt (1992)

	Communalities			
Variable	GR1	GR2	RP1	RP2
PS	.96	.97	.96	.98
IP	.88	.90	.88	.94
IN	.87	.89	.87	.94

	Dimension correlations		
	PS	IP	IN
PS	—		
IP	.79	—	
IN	.92	.85	—

	Exercise correlations			
	GR1	GR2	RP1	RP2
GR1	—			
GR2	.19	—		
RP1	.16	.34	—	
RP2	.15	.31	.43	—

Note. $\chi^2(51, N = 89) = 63.12$, Tucker–Lewis Index = 0.977, relative noncentrality index = 0.98, root-mean-square error of approximation = 0.052. GR1 = first group discussion exercise; GR2 = second group discussion exercise; RP1 = first role-playing exercise; RP2 = second role-playing exercise; PS = problem solving; IP = interpersonal skills; IN = initiative.

(*Appendixes continue*)

Table B4

Correlated Uniqueness Model Parameter Estimates for Combination Model Applied to Multitrait–Multimethod Matrix of Schneider and Schmitt (1992)

	Dimension loadings			Uniquenesses											
				GR1			GR2			RP1			RP2		
Variable	PS	IP	IN	PS	IP	IN	PS	IP	IN	PS	IP	IN	PS	IP	IN
GR1															
PS	.32			.86											
IP		.13		.58	.96										
IN			.40	.61	.51	.78									
GR2															
PS	.44						.80								
IP		.51					.46	.80							
IN			.53				.55	.51	.75						
RP1															
PS	.75									.45					
IP		.67								.33	.57				
IN			.53							.41	.36	.72			
RP2															
PS	.48												.79		
IP		.61											.54	.58	
IN			.49										.64	.50	.77

	Factor correlations		
Factor	PS	IP	IN
PS	—		
IP	.77	—	
IN	.92	.86	—

Note. $\chi^2(39, N = 89) = 40.86$, Tucker–Lewis Index = 0.995, relative noncentrality index = 0.997, root-mean-square error of approximation = 0.026. GR1 = first group discussion exercise; GR2 = second group discussion exercise; RP1 = first role-playing exercise; RP2 = second role-playing exercise; PS = problem solving; IP = interpersonal skills; IN = initiative.

Received June 1, 2000
Revision received January 28, 2001
Accepted February 7, 2001 ■